# Istraživanje o filmovima

Seminarski rad u okviru kursa

Istraživanje podataka

Matematički fakultet

Vojkan Cvijović

vojkancvijovic@gmail.com

18. septembar 2018.

## Sažetak

U svetu postoji sve veće interesovanje za istraživanje, obradu, rukovanje podacima u različite svrhe, stoga je ovo primer rada u kome je vršeno istraživanje sakupljenih podataka o filmovima koji su izašli u periodu između 1889. i 2018. Prikupljeni podaci sadrže osnovne informacije o filmovima poput naslova, žanra, godine snimanja filma, reditelja, glumaca ... Adekvatnim pretprocesiranjem, vizuelizacijom, primenom najpoznatijih algoritama iz oblasti istraživanja podataka prikazani su različiti zanimljivi rezultati, kao rezultat njihove primene.

## Sadržaj

# Uvod

Autor je prikupio podatke tako što je prvo prikupio sve nazive filmova koji su izašli u period između 1989 I 2018. Zatim je koristeći sajt omdbapi.com sakupio ostale potrebne podatke na osnovu imena filmova. Uz pomoć KNIME "*Konstanz Information Miner*" alata, kao I python jezika koji služe za ustraživanje podataka, u radu će biti predstaljeni različiti rezultati o podacima, koji su dobijeni primenom odgovarajućih algoritama pravila pridruživanja, klasterovanja i klasifikacije.

# Podaci

## Opis podataka

Podaci se mogu preuzeti sa linka I smešteni su u pet datoteka. To su "*Movies_Movies.csv*", "*Movies_Genres.csv*", "*Movies_Writer.csv*", "*Movies_Actors.csv*", "*Movies_AdditionalRating.csv*".

U datoteci "*Movies_Movies.csv*" nalaze se uopšteni podaci o filmovima. Datoteka sadrži 178685 redova i 18 kolona. Kolone su:

- Awards, sadrži broj nagrada i nominacija
- Country, država porekla filma
- DVD, datum kada je objavljena DVD verzija filma
- Director, naziv reditelja filma
- Language, najzstupljeniji jezik u filmu
- Plot, kratak opis radnje filma
- Poster, hiperlink filmskog postera
- Production, naziv produkcijske kuće
- Rated, oznaka MPAA ("*Motion Picture Association of America*") kategorije
- Released, datum izlaska filma
- Runtime, vreme trajanja filma
- Title, naslov filma
- Type, tip
- Website, hiperlink do sajta sa informacijama o filmu
- Year, godina izlaska filma
- imdbID, id filma u bazi sajta imdb
- imdbRating, prosečana ocena sa sajta imdb

- imdb Votes, broj ocana filma na sajtu imdb.

U datoteci "*Movies_Genres.csv*" se nalaze žanrovi za filmove iz datoteke "*Movies_Movies.csv*". Datoteka sadrži 308565 redova I 3 kolone. Kolone su:

- Num, označava broj u tableli
- Genre, žanr filma
- imdbID, id filma u imdb bazi, služi da poveže redove iz datoteke "*Movies_Genres.csv*" sa datotekom "*Movies_Movies.csv*"

U datoteci "*Movies_Writer.csv*" se nalaze pisci koji su radili na filmovima iz datoteke "*Movies_Movies.csv*". Datoteka sadrži 66164 redova I 3 kolone. Kolone su:

- Person, ime I prezime osobe
- Responsibility, zaduženje osobe na filmu
- imbdID, id filma u imdb bazi, služi da poveže redove iz datoteke "*Movies_Genres.csv*" sa datotekom "*Movies_Movies.csv*"

U datoteci "*Movies_Actors.csv*" se nalaze glumci iz filmova iz datoteke "*Movies_Movies.csv*". Datoteka sadrži 143869 redova I 3 kolone. Kolone su:

- Num, označava broj u tabeli
- Actors, označava ime glumca
- imdbID, id filma u imdb bazi, služi da poveže redove iz datoteke "*Movies_Actors.csv*" sa datotekom "*Movies_Movies.csv*"

U datoteci "*Movies_AdditionalRating.csv*" se nalaze dodatne ocene filmova iz datoteke "*Movies_Movies.csv*". Datoteka sadrži 92016 redova i 4 kolone. Kolone su:

- Num, označava broj u tabeli
- Rating, ocena filma
- RatingSource, izvor ocene filma
- imdbID, id filma u imdb bazi, služi da poveže redove iz datoteke "*Movies_AdditionalRating.csv*" sa datotekom "*Movies_Movies.csv*"

## Nedostajuće vrednosti

Jedino je u datoteci "*Movies_Movies.csv*" bilo nedostajućih vrednosti. Iz prikaza koliko nedostajućih vrednosti ima, može se primetiti da je najviše takvih vrednosti u vezi sa imdb-om pa tako ubedljivo su prvi imdbRating i imdbVotes. U tabeli ispod prikazani su atributi za koje su nedostajalo podaci.

| Naziv atributa | Originalno | Posle ograničenja |
|---|---|---|
| imdbVotes | 116658 | 80462 |
| imdbRating | 116614 | 80428 |
| Released | 61713 | 44529 |
| Runtime | 59190 | 33282 |
| Language | 25491 | 14730 |
| Director | 27334 | 14469 |
| Country | 12689 | 9107 |
| imdbID | 1 | 0 |
| Title | 1 | 0 |
| Year | 1 | 0 |

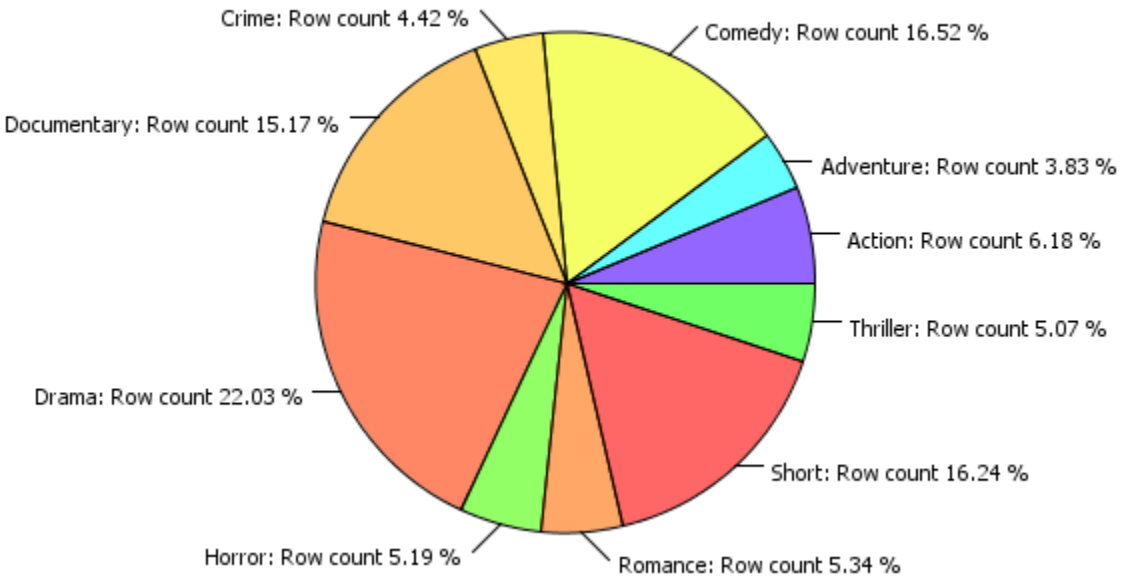
Biramo opseg godina u kojima imamo najviše filmova. Ispod se nalazi tabela brojem filmova za svaku godinu u vremenskom intervalu od 1987. godine do 2018. Uzimamo opseg od 1990 do 2016. Kako ćemo na dalje u radu često koristiti imdbRating, a ukoliko bismo pokušali da umesto brisanja reda koristimo srednju vrednost, ti podaci bi prevladali. Iz tog razloga biramo da brišemo redove koji imaju imdbRating nepoznat. Takodje uklanjamo u red u kome u atribut “Year” nepoznat, kako je to samo jedan red, ne očekuje se da ima uticaja.

| | | | | | |
|---|---|---|---|---|---|
| … | | 1997 | 1799 | 2008 | 6184 |
| 1987 | 953 | 1998 | 2055 | 2009 | 7414 |
| 1988 | 873 | 1999 | 2156 | 2010 | 8160 |
| 1989 | 971 | 2000 | 2497 | 2011 | 8973 |
| 1990 | 1159 | 2001 | 2675 | 2012 | 9815 |
| 1991 | 1175 | 2002 | 3073 | 2013 | 10194 |
| 1992 | 1350 | 2003 | 3573 | 2014 | 11331 |
| 1993 | 1389 | 2004 | 4180 | 2015 | 11687 |
| 1994 | 1494 | 2005 | 4810 | 2016 | 8051 |
| 1995 | 1641 | 2006 | 5285 | 2017 | 3378 |
| 1996 | 1648 | 2007 | 5686 | 2018 | 991 |

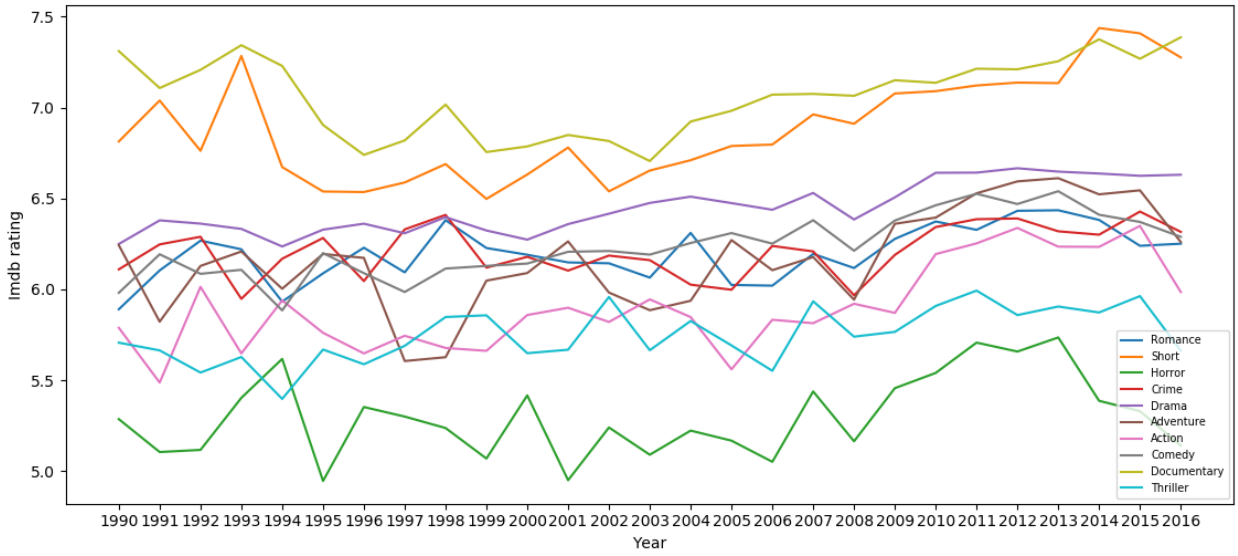
Datoteke “*Movies_Movies.csv*”, “*Movies_Genres.csv*”, “*Movies_Writer.csv*”, “*Movies_Actors.csv*”, “*Movies_AdditionalRating.csv*” su obrađene uz pomoć python skripte obrada.py i kao rezultat dobili smo “*movies.csv*”, “g*enres.csv*”, “w*riter.csv*”, “a*ctors.csv*”, “*additionalRating.csv*”. Iz sve četiri datoteke izbacena je “Unnamed” kolona. Iz datoteke “*Movies_Movies.csv*” izbačeni su atributi 'Awards', 'DVD', 'Plot', 'Poster', 'Production', 'Website', 'Type', 'Rated'.

Razlog za izbacivanje:

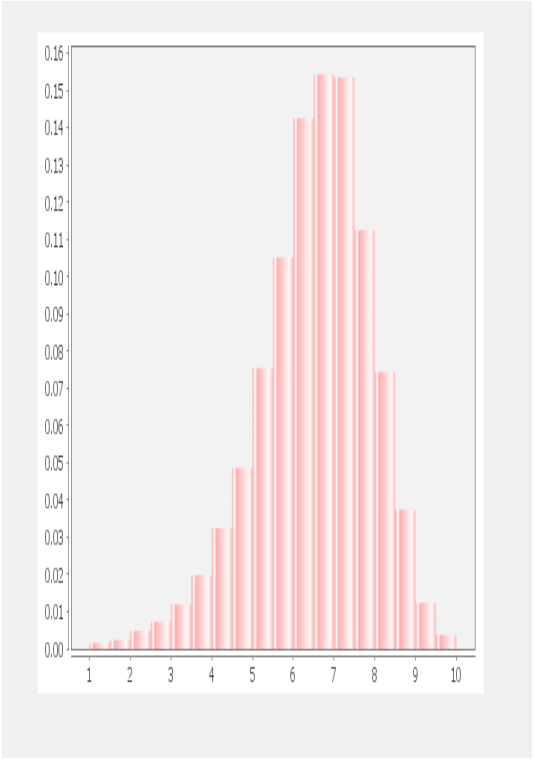
- 'Awards' – 89% slučajeva je prazno, tesko za parsiranje i vrednovanje
- 'DVD' – Godina nam je dovoljan atribut, nismo imali nista posebno vezano za datum DVD izdanja.
- 'Plot' – Nećemo analizirati radnju filma.
- 'Poster' – Nećemo se baviti obradom slika
- ’Production’ – 91% slučajeva je prazno, nije prioritet
- 'Website’ – Ne koristimo u analizama
- 'Type' – Uvek ima vrednost ‘*movie*’
- 'Rated' – Ne spadau atribute koje koristimo za analizu.



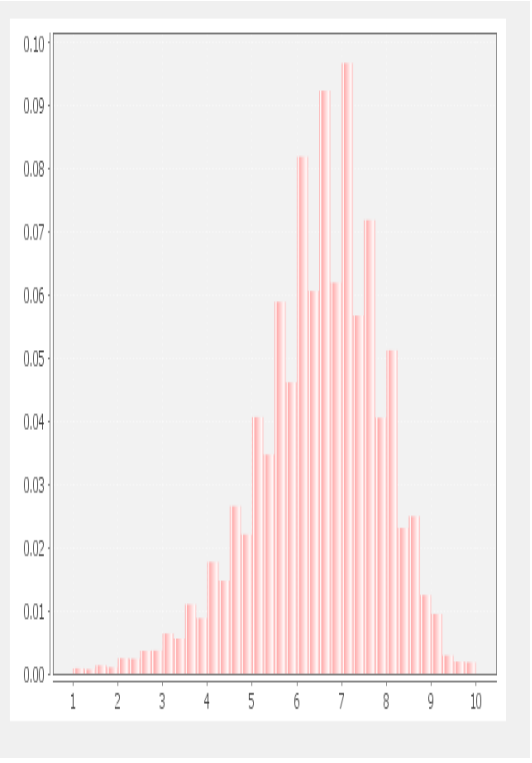

**Slika 1:** Top 10 najzastupljenijih žanrova




Slika 2: Pregled rejtinga po godinama za top 10 žanrova



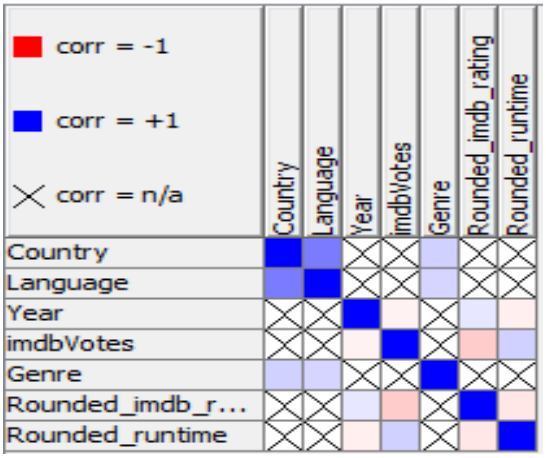
**Slika 3:** Raspodela rejtinga, širina 0.5



**Slika 4:** Raspodela rejtinga, širina 0.25

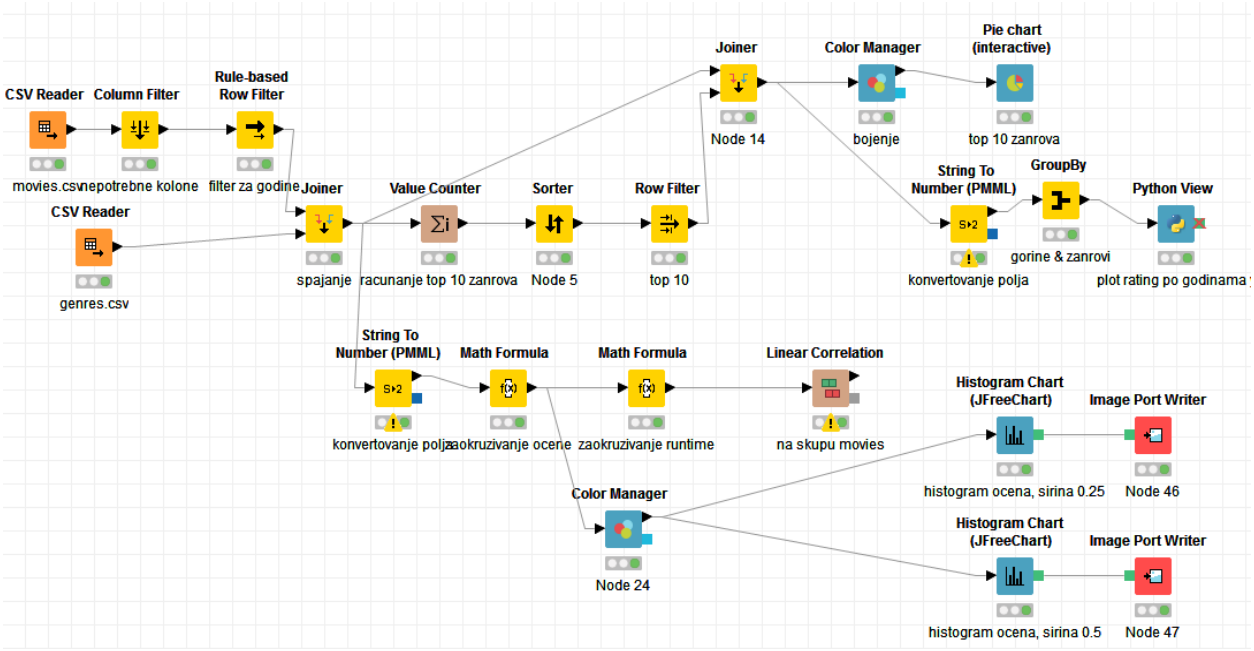
## Korelacija među atributima

Na slici ispod prikazana je matrica linearne korelacije nekih atributa datoteke “m*ovies.csv*”. Najveća korelacija se moze primetiti između jezika i države, što je i očekivano. Dalje postoji slaba korelacija između države i žanra, jezika i žanra. ‘Rounded_runtime’ je vrednost atributa ‘Runtime’ zaokružena na veću cifru deljivu sa pet ( ako je vrednost atributa ‘Runtime’ jednaka 97 vrednost atributa ‘Rounded_runtime’ će biti 100 ). ‘Rounded_imdb_rating’ je vrednost atributa ‘Imdb_rating` zaokružena veću cifru deljivu sa 0.5 ( ako je vrednost atributa ‘Imdb_rating` jednaka 7.7 vrednost atributa ‘Rounded_imdb_rating’ će biti 8 ). Atributi ‘Rounded_imdb_rating’ i ‘Rounded_runtime’ su uvedeni kako bi se smanjili intervali mogućih vrednosti i lakše uočila korelacija među atributima.




**Slika 5: Prikaz korelacije među atributima**

Može se uočiti i negativna korelacija između atributa ‘ImdbVotes’ i ‘Rounded_imdb_rating’, što znači da sa rastom broja glasova opada ocena filma.




**Slika 6: KNIME**

# Pravila pridruživanja

Kako bi se primenila pravila pridruživanja i otkrila neka zanimljiva pravila, neophodno je pretprocesirati podatke. U oblasti Nedostajuće vrednosti opisano je koji su atributi izbačeni kao i redovi sa atributima ‘imdb_rating’ i ‘Years’ kao i razlog. Jedini atribut koji je učestvovao u pravilima pridruživanja a koji je

mogao da ima nedefinisanu vrednost je ‘Directors’ iz datoteke ‘movies.csv’. Iz tog razloga kada god određujemo pravila pridruživanja koja uključuju i atribut ‘Directors’ koristimo čvor ‘Missing Value’ i izbacujemo red sa nedefinisanim poljem. Atribut ‘Genres’ je zaokružen na niži broj deljiv sa 0.5

Neka interesantna pravila pridruživanja:

- Reditelj i imdb rejting

| Support | Confide... | Lift | Conseq... | implies | Items |
|---|---|---|---|---|---|
| 0 | 0.833 | 25.058 | 4.0 | <--- | [Paul Ziller] |
| 0 | 1 | 9 | 7.5 | <--- | [Arthur Ginsberg] |
| 0 | 0.833 | 8.241 | 5.5 | <--- | [Jonathan Liebesm... |
| 0 | 0.875 | 7.875 | 7.5 | <--- | [Mick Thomas] |
| 0 | 0.833 | 7.5 | 7.5 | <--- | [Alfonso CuarÃ³n] |
| 0 | 0.833 | 7.5 | 7.5 | <--- | [Brian Klein] |
| 0 | 0.833 | 5.504 | 6.5 | <--- | [David Mamet] |
| 0 | 0.833 | 5.504 | 6.5 | <--- | [Phil Mulloy] |

- Glumac i imdb rejting

| Row ID | Support | Confide... | Lift | Conseq... | implies | Items |
|---|---|---|---|---|---|---|
| rule0 | 0 | 0.833 | 10.958 | 5.0 | <--- | [ Miuccia Prada] |

- Reditelj i glumac

| Row ID | Support | Confide... | Lift | Conseq... | implies | Items |
|---|---|---|---|---|---|---|
| rule2 | 0 | 1 | 9,557 | Stephan Ma... | <--- | [ Sonja Ball] |
| rule3 | 0 | 1 | 9,557 | Sonja Ball | <--- | [Stephan MartiniÃ¨re, Cassandra Schafhausen] |
| rule5 | 0 | 1 | 7,645.6 | Toshiyuki Hir... | <--- | [ Kathleen Barr] |
| rule6 | 0 | 1 | 7,645.6 | Kathleen Barr | <--- | [Toshiyuki Hiruma, Takashi] |
| rule1 | 0 | 1 | 6,371.333 | Dan Eckman | <--- | [ D.C. Pierson] |
| rule0 | 0 | 1 | 3,822.8 | Max Hardcore | <--- | [ Max Hardcore] |
| rule7 | 0 | 1 | 3,822.8 | Baz Luhrmann | <--- | [ Miuccia Prada] |
| rule4 | 0 | 1 | 509.707 | Kevin Dunn | <--- | [ Steve Austin] |

- Pisac i žanr

| Row ID | Support | Confide... | Lift | Conseq... | implies | Items |
|---|---|---|---|---|---|---|
| rule12 | 0 | 1 | 6.624 | Comedy | <--- | [Adam Herz] |
| rule15 | 0 | 1 | 5.127 | Drama | <--- | [Martha Williamson] |
| rule5 | 0 | 0.75 | 4.968 | Comedy | <--- | [Neil Simon] |
| rule6 | 0 | 0.75 | 4.968 | Comedy | <--- | [Shawn Wayans] |
| rule8 | 0 | 0.7 | 4.637 | Comedy | <--- | [Scot Armstrong] |
| rule10 | 0 | 0.7 | 4.637 | Comedy | <--- | [Marlon Wayans] |
| rule3 | 0 | 0.667 | 4.416 | Comedy | <--- | [Harry Elfont] |
| rule4 | 0 | 0.667 | 4.416 | Comedy | <--- | [Deborah Kaplan] |
| rule7 | 0 | 0.667 | 4.416 | Comedy | <--- | [John Hamburg] |
| rule13 | 0 | 0.667 | 4.416 | Comedy | <--- | [Sacha Baron Cohen] |
| rule14 | 0 | 0.667 | 4.416 | Comedy | <--- | [Todd Phillips] |
| rule9 | 0 | 0.636 | 4.216 | Comedy | <--- | [Dan Mazer] |
| rule11 | 0 | 0.636 | 4.216 | Comedy | <--- | [Ben Stiller] |
| rule2 | 0 | 0.75 | 3.846 | Drama | <--- | [Marc Cholodenko] |
| rule0 | 0 | 0.667 | 3.418 | Drama | <--- | [Cristina Comencini] |
| rule1 | 0 | 0.667 | 3.418 | Drama | <--- | [Sekhar Kammula] |

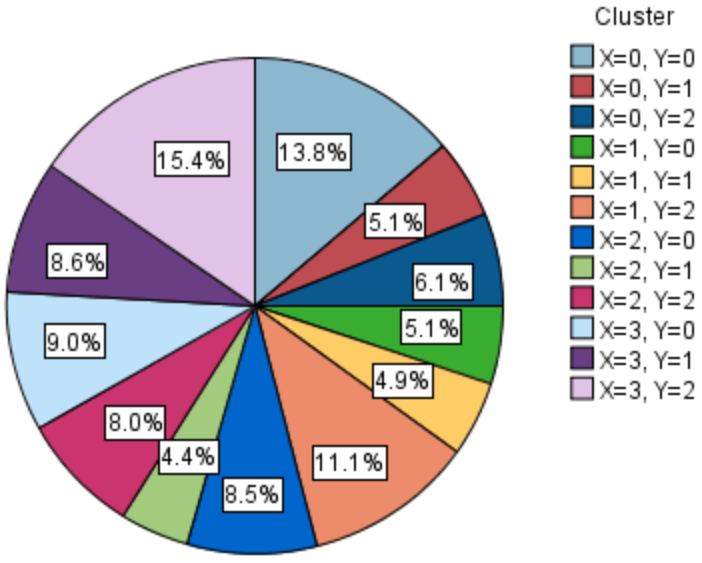
- Reditelj i žanr

| Row ID | Support | Confide... | Lift | Conseq... | implies | Items |
|---|---|---|---|---|---|---|
| rule26 | 0 | 0.812 | 91.386 | Sport | <--- | [Anthony Giordano] |
| rule1 | 0 | 1 | 50.357 | Adult | <--- | [Bunny Luv] |
| rule2 | 0 | 1 | 50.357 | Adult | <--- | [Joey Silvera] |
| rule4 | 0 | 1 | 50.357 | Adult | <--- | [Anthony Spinelli] |
| rule5 | 0 | 1 | 50.357 | Adult | <--- | [Jules Jordan] |
| rule6 | 0 | 1 | 50.357 | Adult | <--- | [Mike John] |
| rule7 | 0 | 1 | 50.357 | Adult | <--- | [Ed Powers] |
| rule8 | 0 | 1 | 50.357 | Adult | <--- | [John Rutherford] |
| rule9 | 0 | 1 | 50.357 | Adult | <--- | [George Duroy] |
| rule12 | 0 | 1 | 50.357 | Adult | <--- | [Chi Chi LaRue] |
| rule13 | 0 | 1 | 50.357 | Adult | <--- | [Paul Norman] |
| rule14 | 0 | 1 | 50.357 | Adult | <--- | [Max Hardcore] |
| rule15 | 0 | 1 | 50.357 | Adult | <--- | [HervÃ© Bodilis] |
| rule16 | 0 | 1 | 50.357 | Adult | <--- | [Rocco Siffredi] |

# Klasterovanje podataka

Ideja klasterovanja jeste da se pronađu grupe objekata, takvih da su objekti u grupi međusobno slični ( ili povezani). Prilikom kalsterovanja podataka primenom K-sredina algoritma, dobijeni su lošiji rezultati, tako da u daljem teksu neće biti razmatrani. Sa obzirom da nema monog kontinualnih atributa u nastavku ćemo razmatrati klastere sa po dva atributa. Klasterovanje sa više atributa, primenom bilo primenom K-sredina algoritma ili Kohonen algoritma, dobijaju se loši modeli sa siluetom 0.1-0.3.

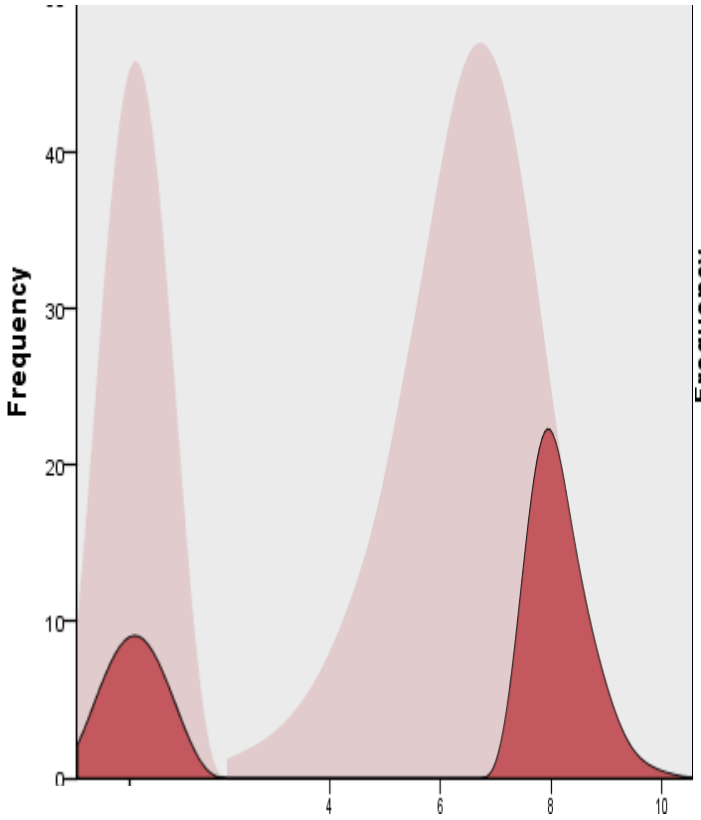
## Rejting i vreme trajanja filma



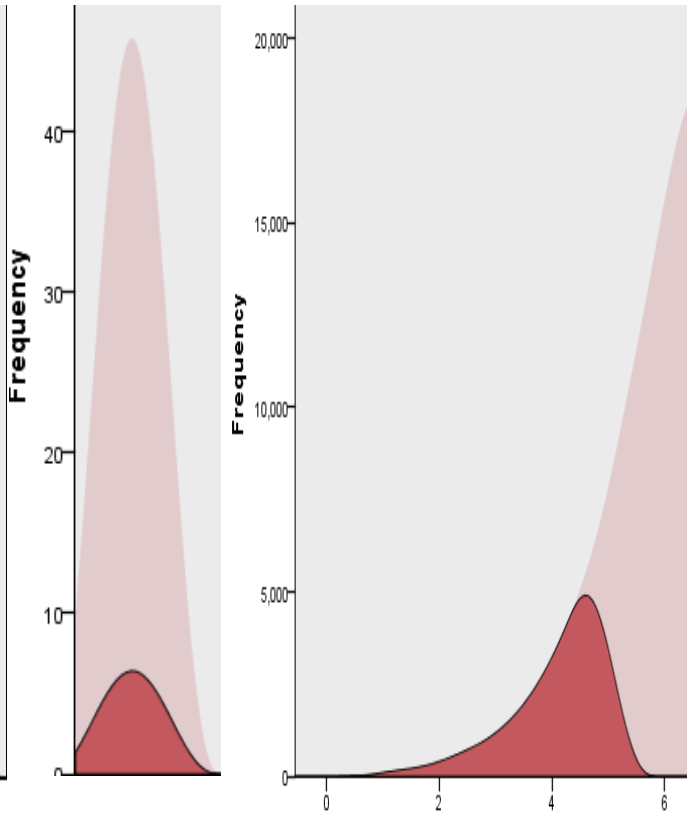

Primenom Kohonen algoritma na skup sa atributima “Imdb_rating” i “Runtime”, dobija se solidan model sa siluetom 0.4. Izdvaja se 12 klastera gde najmanji klaster ima 2623 elemenata (4.4% skupa) a najveći ima 9160 (15.4%). Pa se tako na primer izdvaja klaster (9%) koji obuhvata sve filmove pre 1995 bez obzira na rejting. Takođe imamo i klastere (8.5%) i (11.1%) koji obuhvataju period između 2005-2010 gde prvom pripadaju slabije rangirani filmovi, ispod 6 a drugom bolji iznad 6.

## Vreme trajanja i rejting filma

Primenom Kohonen algoritma na skup sa atributima “Runtime” i “Imdb_rating”, dobija se solidan model sa siluetom 0.5. Izdvaja se 12 klastera od kojih je najmanji klaster is 1686 članova (2.8%) a najveći sa 11769 (19.8%). Izdvajaju  se dva klastera, jedan najbolje ocenjenih filmova, u njemu se nalaze svi filmovi sa ocenom većom od 8 kao i klaster sa najlošijim ocenama, gde su svi filmovi sa ocenama ispod 5, ono što je zanimljivo je da se raspodela vremna trajanja filma ne razlikuje mnogo od ukupne raspodele vremena trajanja filma pa iz toga sledi da vreme trajanja filma nije presudno kada su najbolje i najgore ocene u pitanju.

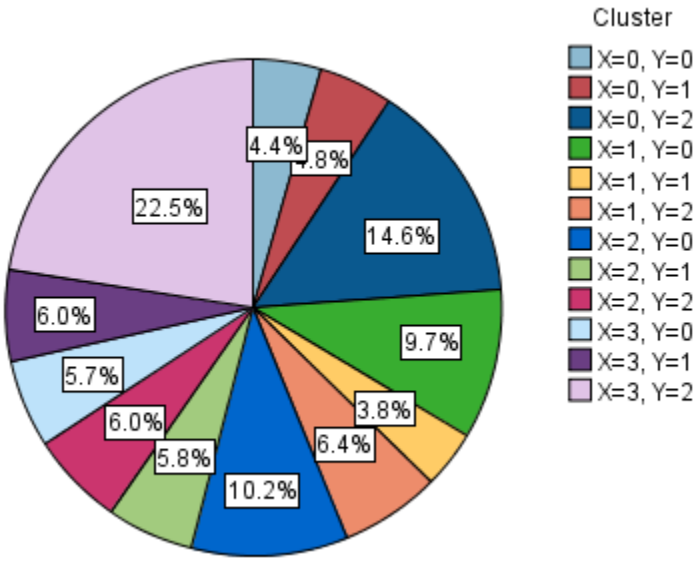



**Frekvencija i rejting najbolje rangiranih filmova**




**Frekvencija i rejting najgore rangiranih filmova**

## Vreme trajanja i godina filma

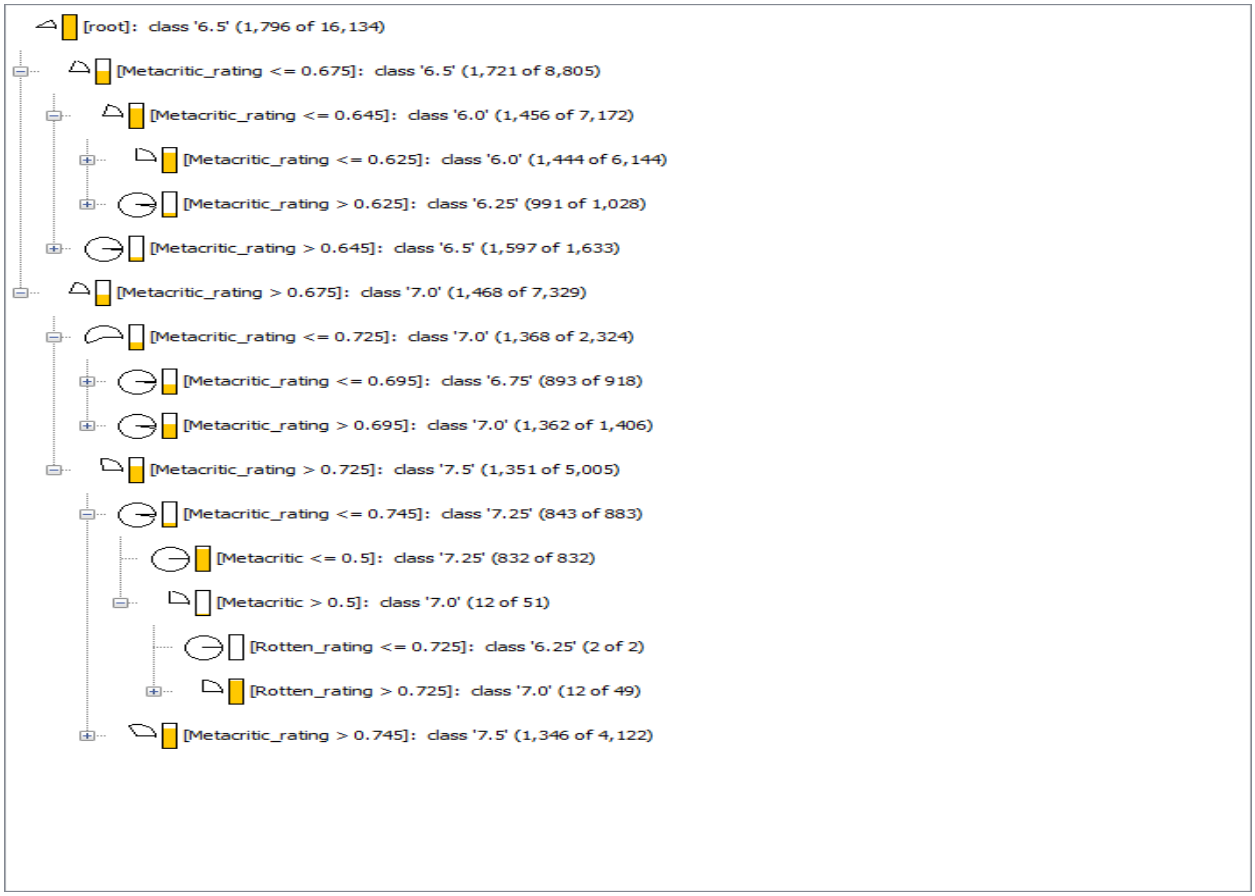



Primenom Kohonen algoritma na skup sa atributima “Runtime” i “Years” dobija se do sada najbolji model sa koeficijentom siluete 0.7. Najmanji klaster ima 2272 (3.8%) filmova dok najveći ima 13313 (22.5%). Izdvajamo dva klastera, jedan kome pripadaju filmovi od 1990 do 2000 a drugi kome propadaju svi filmovi posle 2015. I u jednom i u drugom je slična raspodela dužina trajanja filma. Generalno u svim klasterima je otprilike pojednaka zastupljenost svih dužina trajanja filma.

## Klasifikacija

Klasifikacijom pokusavamo da na osnovu rejtinga iz datoteke “Additional_Ratings.csv” da predpostavimo rejting koji će film imati sa imdb sajta. Koristili smo tri algoritma. Naivni bajesov, k najbližih suseda i drvo odlučivanja. Ocene sa sajta imdb su zaokružene na prvi veći broj deljiv sa 0.25. Podaci su podeljeni 70% za trening 30 % za test. Dobijene su  sledece preciznosti:

- Naive bayes – 0.273
- KNN – 0.855 za k=3
- Tree – 0.921




Ukoliko bi smanjili preciznost na tačnu ocenu, bez zaokruživanja, onda dobijamo sledeće rezulrate:

- Naive bazes – 0.141
- KNN – 0.841 za k = 3
- Tree – 0.679